Kutatók honlapjainak automatikus osztályozása pozitív és jelöletlen tanulás módszerével

Nagy István[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem, Informatikai tanszékcsoport
6720, Szeged, Árpád tér 2.
nistvan@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

1 Bevezetés

Az utóbbi években a kutatók kapcsolatainak feltérképezése és feldolgozása igen intenzíven kutatott területté vált [1]. Egyes kutatók weboldalán számos hasznos életrajzi információ található, úgymint a témavezetők vagy diákok neve, érdeklődési kör, nemzetiség, affiliációk, tudományos fokozatuk stb. [2]. Ezen adatok normalizált változatainak segítségével könnyen feltárható a kutatók közötti kollegiális kapcsolat, (az egy időben és helyen együtt dolgozók) ami nagyban különbözhet az együtt publikálóktól. Mindazonáltal lehetőség nyílik az olyan jellegű kérdések megválaszolására, mint hogy *az amerikai vagy az európai kutatók változtatják gyakrabban a munkahelyüket*.

Az ilyen jellegű feladatok megoldására használt webbányász rendszerek az internet redundanciáját használják ki [3], vagyis azon az elméleten nyugszanak, miszerint minél hasznosabb egy információ, annál többször fordul elő a weben. Ezért használhatóak olyan pontosságra optimalizált algoritmusok, amelyek automatikusan képesek összegyűjteni az adatokat, ugyanakkor nem céljuk az adott információnak az összes elérhető forrásból való kinyerése. Az egyes kutatókról elérhető életrajzi információ sok esetben csak a saját honlapjukon férhető hozzá, ezért, ellentétben a jelenleg alkalmazott megoldásokkal, elengedhetetlen ezen adatok minden esetben való felkutatása [2].

Ebben a cikkben olyan megoldásokat ismertetünk, amelyek automatikusan képesek azonosítani az egyes kutatókhoz tartozó oldalakat. A probléma nehézségét az adja, hogy egy egyszerű webes keresés eredményeként gyakran előfordulhat, hogy a találati lista számos irreleváns oldalt tartalmaz. Ennek egyik lehetséges oka lehet: egy, a keresett kutatóval azonos nevű színész, politikus, esetleg sportoló honlapja kerül a találati listába. Ugyanakkor nehézséget jelenthet az adott kutató által írt könyveket, publikációkat ajánló oldalak kiszűrése is. Ezért az egyes kutatók internetes oldalainak azonosítása érdekében a kereséshez használt online keresők eredményeit automatikusan, „kutató honlap” és „irreleváns honlap” csoportokba kell sorolni. A probléma megoldásához az utóbbi években igen intenzíven kutatott, *pozitív és jelöletlen mintából tanulás* standard módszereit és néhány általunk megkonstruált algoritmust mutatunk be.

2 Kutatók honlapjainak automatikus azonosítása

Amióta az internet különböző információk óriási adatbázisává vált, a honlapok automatikus osztályozása vagy kategorizálása igen intenzíven kutatott terület lett. A probléma megoldására adott legígéretesebb megközelítések a pozitív és jelöletlen tanulás valamely változatát alkalmazták, melyeknek legfőbb előnye a klasszikus osztályozókkal szemben, hogy a tanulás során nincs szükségük negatív példákra.

Az egyes kutatók honlapjainak az azonosítása során (azok kiválasztása a webes keresés találatai közül) 89 kutató letöltött honlapján, annotátorok által előzetesen bejelölt affiliációkat tartalmazó korpuszt használtuk. Amennyiben egy oldal tartalmazott jelölt affiliációt, akkor azt pozitív példának tekintettük, egyébként negatívnak. Az így kialakított dokumentumhalmaz 177 pozitív és 229 negatív példát tartalmazott. Az osztályozáshoz a modell által kialakított nagydimenziós térben is hatékony döntési fákat alkalmaztuk. Ezen megközelítés legnagyobb előnye, hogy az ember számára könnyen értelmezhető outputot generál, ráadásul éppen diszkrét jellemzők feldolgozására fejlesztették ki.

Az adott feladatot hatféleképpen oldottuk meg, melynek eredményeit az első táblázat hivatott illusztrálni. A korpuszt, a szövegbányászati modellek első, és egyben egyik legszélesebb körben használt dokumentum reprezentációs eszközével, a vektortérmodellel illusztráltuk. A különböző megközelítések alapvetően az egyes dokumentumokat leíró vektorokban különböztek. Ennek alapvető oka, hogy megpróbáltuk különböző tartalom alapján elvégezni a honlapok osztályozását. Az első táblázat első sorában egy dokumentumot a hozzátartozó URL és az abból kialakított n-gramok illusztrálják. A második, harmadik és negyedik sorban egy online keresés során elérhető snipet információk segítségével reprezentáltuk a teret. Az utolsó két sorban az adott honlap teljes szöveges tartalma és a hozzá tartozó webcím jelentette a reprezentáció alapját.

1. táblázat: Kutatók honlapjainak osztályozása.

Megközelítés	Pontosság	Fedés	F-mérték
URL	0,785	0,786	0,786
Snipet + URL	0,763	0,764	0,763
Snipet	0,828	0,828	0,826
Snipet + szűrők	0,845	0,845	**0,845**
Honlap + URL	0,79	0,791	0,79
Honlap + URL + szűrők	0,853	0,852	**0,852**

Az első táblázatban jól látható, hogy a keresés során elérhető snipet adatok és a honlapok teljes tartalmát különböző szűrőkkel és az URL-lel kiegészítve sikerült a legjobb eredményt elérni. Ennek megfelelően a későbbiekben ezen megközelítések eredményeit ismertetjük.

A pozitív és jelöletlen példákból való tanuláshoz a fentiekben leírt korpuszt alkalmaztuk. Minden negatív dokumentumot, valamint minden második pozitívot „*jelöletlen*” címkével láttunk el. Az így kialakult dokumentumhalmazt még kiegészítettünk további 30 kutatóhoz tartozó csaknem 200 újonnan letöltött dokumentummal, amik

szintén „*jelöletlen*” címkét kaptak. Ugyanakkor a kiértékelés természetesen az eredeti korpuszon történt.

2. táblázat: Pozitív és jelöletlen tanulás eredményei.

Algoritmus	Pozitív F (honlap)	F (honlap)	Pozitív F (snipet)	F (snipet)
PEBL	0,25	0,68	0,61	0,26
PEBLII	0,62	0,57	0,62	0,57
Tf-idf PEBL	0,65	0,62	0,62	0,57
Rocchio	0,61	0,26	0,61	0,26
Rocchio-Cluster	0,61	0,26	0,61	0,26
Rocchio PEBL	0,60	0,55	0,63	0,57
Spy	0,43	0,69	0,42	0,71
Modosított PEBL	**0,78**	**0,806**	**0,72**	**0,745**
Szavaztatás	**0,76**	**0,769**	**0,82**	**0,837**

A második táblázatban a tanulás pozitív és jelöletlen példákból különböző népszerű [4, 5, 6] és ezek általunk módosított algoritmusainak eredményei láthatók. A második és harmadik oszlopban a honlapok teljes szöveges tartalmából és a hozzájuk tartozó internetcímből képzett n-gramokból álltak az egyes dokumentumokat leíró vektorok, míg a harmadik és negyedik oszlopok csak a keresés során elérhető snipet adatokat tartalmazták.

A pozitív és jelöletlen tanulás egyik első, úttörő algoritmusa a PEBL (más néven 1-DNF vagy M-C) [4]. A megközelítés lényege, hogy a pozitív halmazban leggyakrabban előforduló szavak kigyűjtése után, azokat a dokumentumokat jelöljük negatívnak a jelöletlen halmazból, amelyekben egyetlenegyszer sem fordult elő ezen szavakból. Hátránya, hogy gyakran egyetlen dokumentumot sem jelöl negatívnak (a snipet esetben is így történt). Éppen ezért a PEBLII algoritmusnál [5] könnyítettek a feltételeken. Ebben az eseten akkor kerül be egy szó a pozitív szólistába, ha annak frekvenciája nagyobb a jelöletlen halmazbelinél, ugyanakkor meghalad egy bizonyos értéket. Az általunk kidolgozott tf-idf PEBL esetében, hogy az adott problémára minél inkább jellemző szavak kerüljenek a pozitív szólistába, ezért a mindkét halmaz tf-idf súlyozása után, szintén azok szavak kerülnek kiválogatásra, amelyek frekvenciája magasabb a pozitív halmazon. Mindhárom algoritmus igen hatékonynak bizonyul, amennyiben sikerül a helyes paraméterezést beállítani. A Rocchio algoritmus lényege, hogy az egyes tf-idf súlyok és a halmazok alapján minden csoporthoz egy-egy középpontot határoz meg, és az egyes elemeket ezekhez rendeli. A Rocchio-Cluster az előző megközelítés egy finomítása, miszerint a jelöletlen halmazt összefüggő csoportokra bontjuk, majd minden egyes halmazhoz meghatározzuk a középpontokat. A Spy megközelítés [6] lényege, hogy a pozitív példák egy részét a jelöletlenek közé másoljuk, ezáltal megkönnyítve a jelöletlen halmazban a pozitív dokumentumok „leleplezését”. Az általunk megvalósított módosított PEBL algoritmus lényege, hogy a pozitív szólistába egészen addig kerülnek bele a jellemző szavak, amíg a kezdeti negatív halmaz mérete meg nem egyezik a pozitívéval. A Rocchio PEBL algoritmus negatív középpontját a módosított PEBL megközelítés által kijelölt halmazon számoljuk ki, ezáltal az távolabb kerül a pozitív középponttól. Végül a szavaztatás megköze-

lítés esetében akkor kerül egy elem a kezdeti negatív halmazba, ha a Spy, Rocchio vagy a módosított PEBL algoritmusok közül legalább kettő negatívnak jelölte.

A második táblázatban jól látható, hogy az általunk megvalósított és módosított algoritmusok érték el a legjobb eredményeket. Továbbá a jelen feladat során kiemelten fontos pozitív fedésben is a legjobbak közt teljesítettek.

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND projekt (Jedlik Ányos program) keretében az NKTH támogatta.

Hivatkozások

1. Said, Y. H., Wegman, E. J., Sharabati, W. K., Rigsby, J. T.: Social networks of author-coauthor relationships. Computational Statistics & Data Analysis, 52(4) (2008) 2177–2184
2. Nagy, I., Farkas, R., Jelasity, M.: Researcher affiliation extraction from homepages. In: Proceedings of the NLPIR4DL Workshop at ACL (2009)
3. Califf, M. E., Mooney, R. J.: Relational learning of pattern-match rules for information extraction. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence (1999) 328–334
4. Yu, H., Han, J., Chang, K. C.: PEBL: positive example based learning for Web page classification using SVM. In: KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (2002) 239-248
5. Zuo, W., Yu, H., Peng, T.: A New PU Learning Algorithm for Text Classification A New PU Learning Algorithm for Text Classification. In: MICAI 2005: Advances in Artificial Intelligence (2005) 824-832
6. Li. X., Li, L. B., Ng, S.-K.: Learning to Classify Documents with Only a Small Positive Training Set. In: Machine Learning: ECML 2007 (2007) 201-213